DOCUMENT RESUME

ED 384 663 TM 023 957

AUTHOR Stocking, Martha L.
TITLE Controlling Item Exposure Rates in a Realistic Adaptive Testing Paradigm.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-93-2
PUB DATE Jan 93
NOTE 52p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Adaptive Testing; *Computer Assisted Testing; *Item Banks; Models; Test Format; *Test Items
IDENTIFIERS Paper and Pencil Tests; Probabilistic Models; *Randomization; Test Repeaters

ABSTRACT

In the context of paper and pencil testing, the frequency of the exposure of items is usually controlled through policies that regulate both the reuse of test forms and the frequency with which a candidate may retake the test. In the context of computerized adaptive testing, where item pools are large and expensive to produce and testing can be on a continual basis, new strategies are required. This paper discusses the popular randomization strategy for controlling item security and a less well known probabilistic approach due to Sympson and Hetter. Extensions are developed to the Sympson and Hetter approach to make it more relevant for modern adaptive testing. Examples are given of the application of the randomization approach and the extended Sympson and Hetter approach. (Contains 5 tables, 5 figures, and 11 references.) (Author)

RESEARCH REPORT

CONTROLLING ITEM EXPOSURE RATES IN A REALISTIC ADAPTIVE TESTING PARADIGM

Martha L. Stocking

Educational Testing Service
Princeton, New Jersey
January 1993

CONTROLLING ITEM EXPOSURE RATES IN A REALISTIC ADAPTIVE TESTING PARADIGM

Martha L. Stocking

Educational Testing Service
Princeton, New Jersey 08541

December, 1992

CONTROLLING ITEM EXPOSURE RATES IN A REALISTIC ADAPTIVE TESTING PARADIGM

Abstract

In the context of paper and pencil testing, the frequency of the exposure of items is usually controlled through policies that regulate both the reuse of test forms and the frequency with which a candidate may retake the test. In the context of computerized adaptive testing, where items pools are large and expensive to produce and testing can be on a continual basis, new strategies are required. This paper discusses the popular randomization strategy for controlling item security and a less well known probabilistic approach due to Sympson and Hetter. Extensions are developed to the Sympson and Hetter approach to make it more relevant for modern adaptive testing. Examples are given of the application of the randomization approach and the extended Sympson and Hetter approach.

Key words: computerized adaptive testing, item exposure control, test security, exposure rates.

CONTROLLING ITEM EXPOSURE RATES IN A REALISTIC ADAPTIVE TESTING PARADIGM

Introduction

Every year millions of conventional paper-and-pencil tests are administered by various national testing agencies. These tests are "high stakes" tests in that important decisions about candidates are based, in part, on test scores. Examples include college admissions tests such as the Scholastic Aptitude Tests and the ACT Assessments, graduate and professional school admissions tests such as the Graduate Record Examinations and the Graduate Management Admissions Tests, and licensing examinations such as those sponsored by the National Council of State Boards of Nursing and the National Council of Architectural Registration Boards to aid licensing decisions for nurses and architects respectively.

In secure conventional paper-and-pencil testing for national testing agencies, large numbers of candidates take the same or parallel linear test forms at a few fixed administration dates scheduled throughout some time period. By secure we mean that a great deal of time and effort is spent by the testing agency to insure that no candidates have access to test questions in advance of test administration. In this context, the frequency with which a single item might be seen by a single examinee can be tightly controlled through policies that regulate both the reuse of test forms and the frequency with which candidates may retake a test.

For example, a policy could be established that a form may never be reused. Although this policy is expensive in terms of the cost of item writing and test construction, it does insure that no examinee has (legal) access to items before test administration. Other, perhaps less expensive, variants are possible; for example one could specify that a test form may not

be repeated within a year of its administration and that all candidates who wish to retake an examination must do so within six months. However, in this latter situation, an examinee could always attempt to gain some knowledge about test questions by discussion with other examinees who had previously taken a test form.

Adaptive tests are tests in which items are selected from a large pool of items to be appropriate for the examinee (the test "adapts" to the examinee). All but a few proposed designs have assumed that items would be chosen and administered to examinees on a computer, hence the term "computerized adaptive testing" or CAT. (See Lord (1980) or Wainer, et al. (1990) for a more detailed description of adaptive testing.) In this context the work of insuring secure testing requires a different approach. Adaptive test item pools are typically many times larger than a conventional linear test, and therefore cost more to develop and are more expensive to replace frequently. In addition, testing no longer need be on a few fixed dates but can be virtually continuous. Thus candidates have the opportunity to interview examinees who have recently taken an adaptive test using the same item pool and who therefore might be able to convey information about items they have seen. If adaptive testing is to be a serious competitor to conventional paper-and-pencil testing for secure national testing programs, methods must be developed to restrict item exposure to insure fairness to all candidates, as is currently the case for conventional tests.

This paper briefly describes a previously developed realistic paradigm for adaptive testing and reviews past attempts at controlling the frequency of item use. A new methodology, extending the work of Sympson and Hetter (1985),

is reported and examples are given of the application of this new methodology in realistic adaptive testing.

A Realistic Adaptive Testing Paradigm

A realistic adaptive testing paradigm was developed by Stocking and Swanson (1992). In this paradigm, adaptive tests are constructed by employing a methodology borrowed from the decision sciences that models the behavior of expert test specialists. Test specifications incorporating considerations of content, good test development practices, and statistical properties of items are used by the model to select items for an adaptive test. Test specifications can be put into four categories: 1) constraints on some intrinsic property of an item, 2) constraints on item features in relation to all other candidate items, 3) constraints on item features in relation to a subset of items, and 4) constraints on the statistical properties of items.

The control of intrinsic item features is accomplished through the use of explicit constraints, that is, lower and upper bounds (which may be equal) on the desired number of items which possess a particular feature to be included in an adaptive test. For example, a test assembler may want only one or two items that are antonyms using words that frequently appear in the field of science. The relative importance to the test assembler of these types of constraints may be expressed through the use of differential weighting of such constraints. Constraints on item features in relation to all other candidate items are typically controlled through the use of overlap groups. An overlap group consists of a list of items that may not appear together in the same adaptive test.

Constraints on item features in relation to a subset of items are managed through the idea of conceptual partitions of the item pool that can be described as the block structure of the pool. The implication of a block is that the administration of items from a block may not be interrupted by items not belonging to the same block. Items associated with a single stimulus, such as items based on the same reading passage, are typically considered to be a block. Blocks can also be constructed based on any other item feature of interest such as items which are administered using a single set of directions, or items requiring knowledge of particular content areas.

Constraints on the statistical properties of items are managed through the approach that an item has optimum statistical properties if it has the largest item information function (Lord, 1980, equation 5-9) at the examinee's estimated ability level.

Stocking and Swanson (1992, page 19) summarize this adaptive testing paradigm as follows: the next item administered in an adaptive test is the item that simultaneously

1) is the most informative item at an examinee's estimated ability level, and

2) contributes the most to the satisfaction of all other constraints in addition to the constraint on item information.

At the same time, it is required that the item

3) does not appear in an overlap group containing an item already administered, and

4) is in the current block (if the current item is in a block), starts a new block (if the current item finishes a block), or is in no block.

Controlling Item Exposure

Any scheme that seeks to control the exposure of items employs mechanisms that override the optimal item selection procedure, thus degrading the quality of the adaptive test. Longer tests are therefore required to achieve the level of efficiency obtained when only the optimal item selection procedure governs the choice of the next item, but longer tests may be viewed as a reasonable exchange for greater item and test security.

An adaptive test that employs no control over item exposure might work as follows. Unless other mechanisms are employed to determine how to start an adaptive test, every examinee is administered the same first item. Similarly, each examinee will be administered one of a pair of items as a second item, depending upon whether the first item is correct or incorrect. It is likely that the first few items in an item pool would become public knowledge quickly in any such adaptive test that is administered to more than a few candidates.

Early theoretical investigations of adaptive testing ignored this problem (see, for example, Lord, 1970). Procedures that seek to prevent the overexposure of initial items developed when the prospects of actual implementation became more certain. Lord (1977), Stocking (1987), McBride and Martin (1983), and Weiss (1977) implemented strategies typical of these first attempts. In this approach, the selection of the next item to administer is no longer based solely on the evaluation of items for optimality at the current ability estimate, however optimality may be defined in a particular application. Rather, a group of items is identified that are roughly equal in optimality and the next item is chosen randomly from this group.

A typical approach is to select the first item randomly from a group of five, the second randomly from a group of four, the third randomly from a

group of three, the fourth randomly from a group of two, and the fifth and subsequent items chosen to be optimal (McBride and Martin, 1983). The assumption underlying this approach is that after some number of initial items examinees will be sufficiently differentiated so that subsequent items will vary a great deal. Thus it is sufficient to control the exposure of early items while not controlling later items.

Many variations on this theme are possible, of course, including the possibility of never choosing the next item optimally with certainty, that is, the minimum group size is always two or greater. This latter approach recognizes that in spite of randomization on initial items, examinees with similar abilities may receive many of the same items subsequently unless attempts are made to control the exposure of items administered later in the test.

The goal of this type of strategy is to attempt to insure that examinees, even those with the same or similar ability, receive different adaptive tests. This approach works best in simple situations with large item pools where all or most of the items in the pool have approximately the same statistical properties and where there are no other restrictions on item selection due to content or overlap or any block structure present in the item pool. Its success becomes less predictable in more complex situations. For example, it may not work at all in the context of a small block of items associated with a single reading passage. Suppose that there were four such items in a block from which only two are to be administered to any examinee, and suppose further that one item has already been administered. The randomization scheme cannot specify that the next item be selected from more

than three items of approximately the same optimality because there are only three items remaining in this block.

Not only is this approach problematic in complex but realistic adaptive testing situations, it is also difficult to determine the best sequence of group sizes from which random selection is done by anything other than time consuming trial and error, with no certainty of success.

The Sympson and Hetter Approach

The simple procedure described above attempts to increase item security by indirectly reducing item exposure. Sympson and Hetter (1985) tackle the issue of controlling item exposure directly in a probabilistic fashion.

The procedure distinguishes between the probability P(S) that an item is selected as optimal in an adaptive test for an examinee randomly sampled from a typical group of examinees, and P(A|S), the probability that an item is administered, given that it has been selected. If an item is administered every time it is selected as the optimal item, the item might become overexposed. The procedure seeks to control the overall probability that an item is administered, $P(A) = P(A|S)*P(S)$, and to insure that the maximum value over all P(A)s is less than some value r. This value r is the expected (not observed) maximum rate of item usage.

The conditional probability $P(A|S) = k$ is some fraction that indicates the proportion of the time an item is selected that it should actually be administered. The exposure control parameters, k, one for each item, are determined through a series of simulations (described in a subsequent section) using an already established adaptive test design and simulees drawn from a typical distribution of ability.

Once the exposure control parameters have been established, they are used in the adaptive test as follows:

1) Select the next item for administration.

2) Generate a random number uniformly distributed between 0 and 1.

3) If the random number is less than or equal to the exposure control parameter for the selected item, administer the item.

4) If the random number is greater than the exposure control parameter for the selected item, do not administer the item, and remove it from the pool of remaining items for this examinee. Repeat this procedure for the next-most-optimal item. Continue until an item is found that can be administered.

Many items in an item pool may have exposure control parameters of 1.0, implying that if the item is selected, it is always administered. This tends to happen for items that are appropriate for extreme (high or low ability) examinees since there are not very many of these examinees in the typical distribution of ability. It also tends to happen for items that may be of slightly lower quality than the most optimum items for more typical examinees.

If the adaptive test is of length n, then there must be at least n items in the pool that have exposure control parameters of 1.0. If there were not, then for some examinees there might not be enough items in the pool to administer n of them, that is, a complete adaptive test. In the case where there are not n such items with exposure control parameters of 1.0, Sympson and Hetter suggest the reasonable procedure of sorting the values of the exposure control parameters (including those that are equal to 1.0 already) and setting the n largest to 1.0. This has the effect of increasing the exposure rate for the items that are least popular -- a conservative approach.

Extensions to the Sympson and Hetter Approach

The actual example presented by Sympson and Hetter considered only the statistical properties of items in the selection of the next item for administration. In the realistic adaptive testing paradigm described earlier, many other item features are considered in the definition of optimality and item pools are typically complex structures. Two extensions to the basic methodology are required for this context.

The first extension is a simple analog. If the item pool has a block structure, as described above, then there is a fixed number of items that must be administered from each block before a block can be exited. Instead of insuring that there are n items in the pool with exposure control parameters of 1.0, we must insure that there are n_i items that have exposure control parameters of 1.0 in each of i blocks. This guarantees that there will always be enough items in a block to administer for every examinee.

The second extension is more complex. Real item pools frequently contain sets of items based on some common set of stimulus material, as in items based on the same reading passage or items based on the same table or graph. If there are more items associated with stimulus material than are to be administered to a single examinee, then the exposure rate of the stimulus material itself can be different from the exposure rate for any item associated with that stimulus. Consider a stimulus and four associated items, of which only two are to be administered. Which two items are selected for an examinee depends upon all previous responses and the extent to which the two items satisfy the content and overlap constraints on item selection for this particular examinee. Regardless of which two items are selected, the stimulus material is considered to be exposed only a single time to this examinee.

If there are many such sets, and if the quality of items associated with each set varies widely, the Sympson and Hetter methodology controls the rate at which items are exposed, but not the rate at which the stimuli themselves are exposed. This can result in overexposure for some stimuli.

A natural extension to the Sympson and Hetter approach is to apply the same logic to develop exposure control parameters for the stimulus material in addition to items associated with each stimuli. This is done in a manner exactly analogous to that described above, similar to considering the stimuli to form a separate pool. In the operation of the adaptive test, once the exposure control parameters for stimuli and associated items have been established, they are used as follows:

1) Select the next item for administration.

2) If the item is not associated with (new) stimulus material, proceed with the approach outlined by Sympson and Hetter for discrete items. If the item is associated with (new) stimulus material, then proceed to step 3) below.

3) Generate two random numbers uniformly distributed between 0 and 1, one to be associated with the stimulus and the other to be associated with the item.

4) If the first random number is less than or equal to the exposure control parameter for the stimulus, and the second random number is less than or equal to the exposure control parameter for the item, administer the stimulus and the item.

5) If either or both random numbers are greater than the relevant exposure parameters, do not administer the stimulus and item, and

remove the entire set of items from the pool of remaining items for this examinee.

If the constraints for the adaptive test specify that at least m stimuli must be administered, then there must be at least m stimuli in the pool that have exposure control parameters of 1.0. If there were not, then for some examinees there might not be enough stimuli in the pool to administer m of them. In the case where there are not m such stimuli with exposure control parameters of 1.0, the analogous procedure is used to set the m largest to 1.0.

Estimating Exposure Control Parameters

Sympson and Hetter outline a procedure for estimating values of the exposure control parameters for each item in the pool using successive Monte Carlo simulations in a way that guarantees that no item is administered to more than the fraction r of examinees. The extent to which this guarantee holds in practice depends, or course, on the extent to which the estimated statistical properties of items are good approximations to true statistical properties, as well as the extent to which the estimated distribution of true ability approximates the actual distribution of true ability in the examinee population of interest. This procedure, extended to include consideration of stimuli, is as follows:

1) Specify the adaptive testing design, that is, the item pool, the item selection algorithm, the scoring method, the termination rule, and so forth. Also specify the distribution of true ability of interest and the desirable expected maximum exposure rate of any item or stimulus, r.

2) Draw a large ($\geq$ 1000) random sample from the distribution of true ability. Initially assume that all exposure control parameters are equal to 1.0.

3) Simulate the administration of the adaptive test to the sample, and separately record P(S) and P(A) for every item and stimulus. Note the maximum value of P(A).

4) Given the value of r and the observed values of P(S) in the preceding simulation, redefine new k_i as follows:

   If $P_i(S) > r$, then new $k_i = r + P_i(S)$.

   If $P_i(S) \leq r$, then new $k_i = 1.0$.

5) Given the new values of k_i, repeat step 3 and 4 until the maximum observed $P_i(A)$ approaches a limit slightly above r and subsequent simulations are characterized by small oscillations around this value.

6) Administer real adaptive tests using the k_i obtained from the final simulation.

The procedure seeks to insure that $k_i * P_i(S) \leq r$ for all items and stimuli. Thus, in step 4, if $P_i(S)$ is found to be greater than r, k_i must be adjusted to $r \div P_i(s)$ to make the inequality true. If $P_i(S)$ is already less than r, then k_i can equal 1.0 and the inequality remains true.

Examples

Example 1: A Quantitative Adaptive Test With a Large Item Pool

A 25-item fixed length adaptive test design was established for a pool of items designed to measure quantitative ability. This pool contained 518

entries, of which 22 were stimuli such as tables or graphs with a varying number of questions associated with each stimuli. Items and stimuli were classified on 25 different features thought to be important to the measure, and limits were placed on the numbers of items with each feature that could be included in the adaptive test selected for each examinee. There were 88 groups of items identified as overlapping, involving 318 items in the pool. The Stocking and Swanson (1992) adaptive testing paradigm was used for item selection. The test was scored by transforming the final maximum likelihood ability estimate to the number right true score metric of a reference set of 60 items (Lord, 1980, equation 4-9). This latter metric runs from a chance level of 8 to a perfect score of 60.

Table 1 displays the features of interest, the number of entries in the item pool that were identified as having a particular feature, the lower and upper permissible bounds on such features in a 25-item adaptive test, and relative weight given each constraint reflecting the importance of that constraint in the item selection algorithm. For example, the first feature identifies the stimuli of sets of items described as Data Interpretation sets and specifies that there must be two such sets administered in a 25-item adaptive test and that there are 22 such entries in the 518-item pool. Although not indicated in the table, two items will be administered for each set; the range of items for each set in the pool is from four items to 10 items. As another example, the feature numbered 9 identifies items that are classified as Type 4, indicates that 8 such items must be administered in a 25-item adaptive test, and that there are 246 entries in the item pool that have this property. The number of entries in the sixth column of the table

sums to considerably more that the number of entries in the item pool (518) because features are not mutually exclusive.

Two exposure control methods were tried with this adaptive test design. One was a randomization method in which the first item was randomly selected from a group of eight items identified as the best items -- best in the Stocking and Swanson sense of satisfaction of content, overlap, and statistical properties. The second item was selected from a group of seven such items. The group size for the selection of the third item was six, for the fourth item it was five, and so forth. Items were selected optimally from the eighth item to the end of the test.

The other exposure control method tried was the extended Sympson-Hetter (ESH) method described earlier with a desirable expected maximum exposure rate r of .20. Simulations were performed for 100 simulees at each of 11 values on the reported score metric ranging from just above the chance level to just below a perfect score. To perform the unconditional evaluations required for the extended Sympson-Hetter methodology, the item parameters and item responses from a group of over 4000 real examinees who responded to the 60 reference items as an intact test were used to compute an estimated distribution of true ability using the method of Mislevy (1984).

Figure 1 displays the results of 16 iterations of the extended Sympson-Hetter methodology as well as the randomization method in terms of maximum observed exposure rates for discrete items (that is, items not associated with stimuli), stimuli, and items associated with stimuli. The results for the random method are plotted at the final extended Sympson-Hetter iteration.

The first ESH iteration is equivalent to unrestricted optimum item selection because all of the exposure control parameters are initially set to

1.0. In this situation, a discrete item is administered to all examinees as the starting item, thus the observed exposure rate for this item is 1.0. The maximum observed exposure rate for stimuli is .61, and the maximum observed exposure rate for items associated with stimuli is .54. The estimated reliability of this test, computed by the method suggested by Green, et al, (1984, equation 6) is .93.

At the end of 16 ESH iterations, the observed maximum exposure rate for discrete items is .23, for stimuli it is .18, and for items associated with stimuli it is .17. Items in adaptive tests constructed for real examinees using the exposure control parameters for the 16th iteration are expected to have similar exposure rates. The average exposure rate over all items and stimuli is .10 with a standard deviation of .07. The estimated reliability of this adaptive test is .92. For the randomization method, the observed maximum exposure rate for discrete items is .66, for stimuli it is .59, and for items associated with stimuli it is .51. The overall average exposure rate is .13 with a standard deviation of .14. This test had an estimated reliability of .94. It is highly likely that through trial and error one could discover a randomization scheme that produced better exposure rates than the one tried here. Clearly the ESH method produces more acceptable exposure rates than this particular randomization method, without much sacrifice in estimated test reliability.

It is also important to compare exposure control methods in terms of their affects on constraint satisfaction. Table 2 compares the content constraint violations for the two methods. For each constraint the percent of a typical population observed to have violations of that constraint is given, along with the average number of items administered for that constraint. For

example, if the randomization method is used, for the item feature numbered 12, QC 1 Type 4, 12.7 percent of a typical population receive adaptive tests that violated the constraint that between zero and one items with this feature should be administered. On average, there were .91 such items administered across all adaptive tests, indicating that some simulees received adaptive tests with more than one such item. If the exposure control parameters are from the final ESH iteration, 7.1 percent of a typical population have such constraint violations, and the average number of such items across all adaptive tests is .75, indicating that fewer examinees received more than one such item. In general, the extended Sympson-Hetter exposure control approach produced more constraint violations, although all of the violations for both methods occurred only on constraints with low weights and were considered to be minor.

Example 2: A Quantitative Adaptive Test With a Smaller Item Pool

In the 16th extended Sympson-Hetter iteration, there were 243 entries in the item pool that were not used, including three stimuli and their associated items. Given the probabilistic nature of this approach, there is, of course, no guarantee that any of these would not be used in real adaptive testing, or even in a subsequent simulation with the same exposure control parameters but a different random number seed. Nevertheless, it seems extravagant to continue to include these entries in the item pool, given the requirements for storing all the information about these entries including text and graphics, and given the quality of the adaptive tests produced that make no use of these items.

These 243 entries were deleted from the pool, leaving 331 remaining entries. The numbers of items with each feature in this smaller pool are identified in the seventh column of Table 1. Thirteen more iterations of the extended Sympson-Hetter procedure were then performed, using the same adaptive test design and typical distribution of ability, beginning with an iteration using the exposure control parameters that were used in the sixteenth iteration with the larger pool. Additional iterations were not necessary because the procedure had clearly converged. The results of these iterations are shown in Figure 2.

It is clear from this Figure that the efficacy of exposure control parameters are dependent upon the structure of a particular item pool and do not carry over to a new item pool. For the first iteration with the smaller item pool, the maximum observed exposure rate for discrete items is .66, for stimuli it is .17, and for items associated with stimuli it is also .17. These contrast with values from the final iteration with the larger item pool, where the comparable quantities were .23, .18, and .17.

The big difference in the observed maximum exposure rate for discrete items is due to the fact that what determines the exposure control parameter for a particular item is not only the characteristics of a particular item, but how those characteristics compare to those of all other items in the pool. If the pool size is decreased, even if it is decreased by discarding items that were never used in a particular simulation, the selection rate of most moderate to good items increases. Therefore the exposure control parameters for these items must be reduced to insure that their maximum administration rate does not exceed (in expectation) the value of specified for r.

This same phenomenon could, in theory, occur with stimuli and items associated with stimuli. It did not happen in this particular illustration for at least two reasons. First, the number of stimuli removed from the larger pool (three out of 22) was a much smaller proportion than the number of items removed, therefore stimuli selection rates were less affected. Second, the adaptive test design specifies that item sets cannot appear first in the adaptive test, and it is items appearing towards the beginning of an adaptive test that are more likely to have high selection rates.

At the thirteenth ESH iteration with the smaller pool, the maximum observed exposure rate for discrete items is .23, for stimuli it is .19, and for items associated with stimuli it is also .19. The estimated reliability for this adaptive test is .92. The content constraint violations for this simulation are shown in Table 3, and may be compared with those for the larger pool in Table 2. The observed maximum exposure rates, the estimated reliability, and the content constraint violations compare favorably with those from the larger pool, an indication that proceeding to administer adaptive tests from the smaller pool to real examinees is reasonable.

A comparison of the conditional standard errors of measurement (CSEM) for both exposure control methods and for both pool sizes is shown in Figure 3. The smooth curve in this Figure is the conditional standard error of measurement for the reference set of 60 items considered as a conventional test and scored with an estimated number right true score. There is little difference between the conditional standard errors of measurement from the larger and smaller pools using the extended Sympson-Hetter approach to controlling item exposure. There is more difference between the randomization

approach and the extended Sympson-Hetter approach on the larger pool, particularly in the middle range of test scores.

It makes sense that the extended Sympson-Hetter approach has larger conditional standard errors of measurement in the middle score ranges than the randomization approach, and about the same CSEM at the extremes of the score range. In the randomization method, the frequency of administration of optimal items is independent of the distribution of examinee ability, and depends only in the serial item position within the adaptive test. Thus suboptimal item selection is spread fairly evenly throughout the score range.

In contrast, for the extended Sympson-Hetter approach, the frequency of administration of optimal items depends directly on the distribution of examinee ability. The most popular items in the pool, in the sense of being selected for administration with the highest frequencies, are those items that are most appropriate for administration to typical simulees. Therefore it is those items that will have exposure control parameters that are substantially less than 1.0, reducing, sometimes significantly, their frequency of administration even though they have been selected. For these typical simulees, then, items that are less then optimal will be used with greater frequency than one would expect if there were no controls, thus leading to larger conditional standard errors of measurement. The same result does not hold in regions of the score range with very small frequencies of examinees; since there are so few examinees, optimal items for these examinees are administered about as frequently as they are selected.

Example 3: A Verbal Adaptive Test With an Inadequate Item Pool

The previous two examples used an item pool of the same basic structure, as shown in Table 1. Verbal pools tend to have different, more complex, structures. For this example, a 27-item fixed length adaptive test design was established for a pool of items containing 407 entries, of which 24 were stimuli, in this case reading passages. Items and stimuli were classified on 35 different features and limits were placed on the number of items or stimuli that were desired in the adaptive test. There were 291 groups of items identified as overlapping. As before, each stimulus and its associated items is considered to be a block. However, the block structure for this pool is more elaborate than before in that each major item type, Sentence Completion (SNCP), Analogies, and Antonyms (ANTM) also forms a separate block. Again the Stocking and Swanson adaptive testing algorithm was used for item selection, and the test was scored by transforming the final maximum likelihood ability estimate to the number right true score metric of a reference set of 76 items. This metric runs from a chance score of 13 to a perfect score of 76.

Table 4 displays the features of interest, the number of entries in the pool that were identified as having a particular feature, and the bounds and the weights for each constraint used in the item selection algorithm. There is no information readily apparent in this table that would lead to the conclusion that the item pool is inadequate, and yet it proved to be so when the attempt was made to develop exposure control parameters using the extended Sympson-Hetter approach.

The simulations were performed with 100 simulees at each of 13 values on the reported score metric ranging from just above chance level to just below a perfect score. The desired expected maximum exposure rate was set at .20. To

perform the unconditional evaluations required for the extended Sympson-Hetter methodology, the item parameters and item responses from a group of over 5000 real examinees who responded to the 76 reference items were used to compute an estimated distribution of true ability, again using the method of Mislevy (1984).

Figure 4 displays the results of eight iterations of the extended Sympson-Hetter approach. The exposure control parameters for the discrete items have converged by the eighth iteration to produce an observed maximum exposure rate of .24. However, the exposure control parameters for the stimuli and associated items are changing, from about the fifth iteration on, in such a way as to gradually increase the observed maximum exposure rate. By the eighth iteration, the observed maximum exposure rate for stimuli is .54, and for associated items it is .46. This trend is clearly not desirable.

There are two problems with this pool. First, it is very important for every adaptive test to contain exactly two Type 2 reading passages (constraint number 2), but there are only 11 such passages in the pool. Because there are so few passages of this type, from the sixth iteration on, the two least popular passages have had their exposure control parameters artificially set to 1.0 to insure that complete adaptive tests can be administered in the subsequent iteration. In the next iteration, those with exposure control parameters of 1.0 are administered if selected, increasing their exposure. When new exposure control parameters are computed at the end of an iteration, these passages will be considered too popular and their exposure control parameters reduced, while the two least popular passages in that iteration will have their parameters artificially changed to 1.0. Thus exactly which two passages are least popular varies from iteration to iteration, and the

procedure never converges. It is reasonable to suppose that an increase in the number of Type 2 passages in the pool would mitigate this problem.

The pool is also inadequate in terms of the satisfaction of content constraints, as shown in Table 5. Two Type 3 passages were administered to over 30 percent of a typical population, when the desired number was one Type 3 passage. Also two Type 6 passages were administered to about thirty-eight percent of a typical population when only one was desired. This is unacceptable from the test specialists' viewpoint.

This problem occurs because there is an interaction between the desire to have three passages per simulee -- one of Type 1 and two of Type 2 -- and the other constraints on the stimulus material. Each simulee is also supposed to receive exactly one Type 3 passage and exactly one Type 6 passage. These two categories happen to be mutually exclusive, that is, a passage cannot be categorized as being both Type 3 and Type 6. Looking at the passages in the pool in detail, we find the following:

	Type 3	Type 6	Neither	Total
Type 1	7	4	2	13
Type 2	3	5	3	11
Total	10	9	5	24

If each adaptive test must have one Type 3 and one Type 6 passage, and must also have three passages, then it follows that each simulee must also have one passage that is neither Type 3 nor Type 6. But there are only five such passages in the entire pool, and this is far too few if we want a realistic limit on exposure. As a consequence, the constraints on Type 3 and Type 6 passages are frequently violated. The pool needs to be enriched to provide more passages that are neither Type 3 nor Type 6.

Example 4: A Verbal Adaptive Test With an Adequate Item Pool

The inadequate verbal pool was augmented with additional passages, with a particular focus on adding Type 2 passages that were not classified as either Type 3 or Type 6. A pool of 574 entries was initially obtained. The block structure of this augmented pool was simplified to include only reading comprehension passages and their associated items and a 27-item fixed length adaptive test design was established for the augmented pool. Eight extended Sympson-Hetter iterations were conducted which showed none of the convergence problems seen in Figure 4. This pool was then reduced, based on item usage, to a pool of 372 entries for the final extended Sympson-Hetter iterations.

The number of entries matching each classification of stimulus or item for this final reduced pool of 372 entries is shown in the seventh column of Table 4. Even though the total pool is smaller in size than the inadequate pool, it conforms better to the desired content constraints. There are eight more Type 2 reading passages for a total of 32 passages compared to only 24 for the inadequate pool. This is also seen in the following table showing the interrelation of characteristics of reading passages:

	Type 3	Type 6	Neither	Total
Type 1	6	5	2	13
Type 2	5	5	9	19
Total	11	10	11	32

Over half of the additional passages are not classified as either Type 3 or Type 6, and over half of the additional passages are classified as Type 2.

As before, the simulations were performed with 100 simulees at each of 13 values on the reported score metric ranging from just above chance level to just below a perfect score. The desired expected maximum exposure rate was

set at .20. The starting values for the exposure control parameters were the final values obtained from the larger adequate pool, identical to the procedure followed for Example 2.

Figure 5 displays the results of eight iterations of the extended Sympson-Hetter approach. The exposure control parameters for the discrete items have converged by the eighth iteration to produce an observed maximum exposure rate of .24. For stimuli as well as associated items, this number is .19. The mean exposure rate is .10 with a standard deviation of .08. The reliability of this adaptive test is .90.

Table 5 compares the content constraint violations for those constraints with nonzero weights for the inadequate and adequate pools. The large percentages of typical population that received more than one passage classified as either Type 3 or Type 6 have complete disappeared. The remaining constraint violations were judged satisfactory for the adequate pool by test specialists. Based on these two examples, it seems that pool size is less important in producing adequate adaptive tests than conformity to desired test properties.

Discussion

Computerized adaptive testing presents new challenges to item and test security since such tests can be administered on virtually a continual basis. Because of the size and costs of adaptive test item pools, it is unlikely that it is possible to have more than a few item pools in operational use at the same time. In a more realistic approach, the protection of item security assumes the form of suboptimal item selection within a single item pool to decrease the frequency of use of the best items from the pool.

Previously published methods for controlling the frequency of item administration sought to accomplish this by randomly selecting an item for administration from a group of items of approximately equivalent optimality. This method controls exposure rates only indirectly, and best choices of group sizes can only be determined by tedious trial and error approaches. The Sympson and Hetter approach develops exposure control parameters for each item that directly control the frequency of item administration, given the item's selection as an optimal item, in reference to a typical distribution of examinee ability.

Two extensions to the Sympson-Hetter method, developed in this paper, are required to make this approach completely applicable to modern adaptive testing item selection paradigms such as that developed by Stocking and Swanson. First, if the pool contains a block structure, the exposure control parameters must insure that there are sufficient items that can be administered from each block. Second, and more difficult, in adaptive test designs where a subset of items associated with stimuli are to be administered to an examinee, the Sympson-Hetter approach must be extended to simultaneously control the exposure of stimuli themselves as well as the items associated with the stimuli.

The examples demonstrate a number of salient features of the extended Sympson-Hetter approach:

1) ESH is likely to produce lower exposure of items than reasonably constructed randomization approaches, and in a more straight forward manner.

2) Because ESH is developed in reference to a specific assumed population distribution of ability, it is likely to cause more suboptimal item selection for the more numerous typical examinees than for the less numerous

extreme examinees. In situations where there is a fixed criterion conditional standard error of measurement that must be met for all examinees, it is likely to require increased adaptive test length over the randomization method.

3) The success of the ESH approach is dependent upon a specific pool of items. If a pool is augmented or reduced, exposure control parameters must be redeveloped.

4) ESH may diverge if an item pool is inadequate with respect to desired test properties as represented by the constraints on test content.

The probabilistic strategy of controlling exposure rates exemplified by the extended Sympson-Hetter approach worked well on both a large and a smaller item pool for adaptive testing designed to measure quantitative ability. It also worked well on a relatively small but adequate verbal pool with a very different test structure. Based on the results reported here, it seems likely that it will work well in practice for adaptive testing from these item pools and live examinees. Before it can be universally recommended, experience with actual item exposure rates obtained from real, as opposed to simulated, adaptive testing on these and other pools is strongly desirable.

In addition, it seems clear that the extended Sympson and Hetter approach to controlling item exposure in adaptive testing does not have all the features one might eventually require in operational adaptive testing. For example, although the overall exposure rate of an item is controlled, its exposure conditional on ability is not. Thus an item may be exposed to nearly all examinees of a particular ability, even though its overall exposure rate is low. If this is identified as a problem, it may be necessary to develop methods that control conditional item exposure. Also, although exposure rate is controlled across a distribution of ability, it is not controlled across

candidate volume. An item with an exposure rate of .1 will only be seen by approximately 10% of test takers, but if there are a million test takers, the absolute exposure will be quite high. Further research clearly remains to be done in this area if adaptive testing is to become a secure alternative to conventional paper-and-pencil testing.

References

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. Journal of Educational Measurement, 21, 347-360.

Lord, F. M. (1970). Some test theory for tailored testing. In W. H. Holtzman (Ed.), Computer assisted instruction, testing, and guidance. New York: Harper and Row.

Lord, F. M. (1977). A broad-range tailored test of verbal ability. Applied Psychological Measurement, 1, 95-100.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

McBride, J. R., and Martin, J. T. (1983) Reliability and validity of adaptive ability tests in a military setting. In D. J. Weiss (Ed.) New Horizons in Testing (pp223-236). New York: Academic Press.

Mislevy, R. J. (1984). Estimating latent distributions. Psychometrika, 49, 359-381.

Stocking, M. L. (1987) Two simulated feasibility studies in computerized adaptive testing. Applied Psychology: An International Review, 36, 3/4, 263-277, 1987.

Stocking, M. L., and Swanson, L. (1992) A Method for Severely Constrained Item Selection in Adaptive Testing. (Research Report 92-37). Princeton, NJ: Educational Testing Service. (Conditionally accepted for publication.)

Sympson, J. B., and Hetter, R. D. (1985, October). Controlling item-exposure rates in computerized adaptive testing. Proceedings of the

27th annual meeting of the Military Testing Association (pp. 973-977). San Diego, CA: Navy Personnel Research and Development Center.

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, R. J.,Steinberg, L., and Thissen, D. (1990). Computerized Adaptive Testing: A Primer. Hillsdale, NJ: Lawrence Erlbaum Associates.

Weiss, D. J. (Ed.) (1978). Proceedings of the 1977 Computerized Adaptive Testing Conference. Minneapolis: University of Minnesota.

Table 1: Content Constraints and Weights for the Adaptive Quantitative Test

Number	Description	LB[1]	UB[2]	W[3]	N(518)[4]	N(331)[5]
1	Data Interp Set	2	2	11	22	19
2	QC 1	5	5	10	62	39
3	QC 2	4	4	10	73	34
4	QC 3	4	4	10	69	34
5	PS 1	3	3	10	53	26
6	PS 2	3	3	10	44	26
7	PS 3	2	2	10	41	18
8	DI 1	4	4	1	154	135
9	Type 4	8	8	1	246	178
10	QC Type 4	2	2	10	36	20
11	PS Type 4	2	2	10	56	23
12	QC 1 Type 4	0	1	1	12	6
13	QC 2 Type 4	0	1	1	11	6
14	QC 3 Type 4	0	1	1	13	8
15	PS 1 Type 4	0	1	1	28	9
16	PS 2 Type 4	0	1	1	16	9
17	PS 3 Type 4	0	1	1	12	5
18	Type 5	0	1	1	11	8
19	Type 6	0	1	1	8	5
20	QC Type 7	1	10	1	53	27
21	QC Type 8	1	10	1	58	27
22	QC Type 9	1	10	1	54	28
23	QC Type 10	1	10	1	39	25
24	Type 11	1	12	1	63	49

[1]Lower Bound; [2]Upper Bound; [3]Weight;
[4]Number in larger pool; [5]Number in smaller pool

Table 2: Content Constraint Violations for the Adaptive Quantitative Test, Large Pool, Two Exposure Control Methods

Number	Description	LB[1]	UB[2]	W[3]	N[4]	Randomization, Per cent[5]	Randomization, Ave n[6]	ESH, Per cent[5]	ESH, Ave n[6]
1	Data Interp Set	2	2	11	22				
2	QC 1	5	5	10	62				
3	QC 2	4	4	10	73				
4	QC 3	4	4	10	69				
5	PS 1	3	3	10	53				
6	PS 2	3	3	10	44				
7	PS 3	2	2	10	41				
8	DI 1	4	4	1	154				
9	Type 4	8	8	1	246				
10	QC Type 4	2	2	10	36				
11	PS Type 4	2	2	10	56				
12	QC 1 Type 4	0	1	1	12	12.7	.91	7.1	.75
13	QC 2 Type 4	0	1	1	11	1.3	.59	1.0	.52
14	QC 3 Type 4	0	1	1	13	1.0	.50	4.2	.72
15	PS 1 Type 4	0	1	1	28	.2	.65	1.8	.54
16	PS 2 Type 4	0	1	1	16	16.5	1.10	17.5	1.00
17	PS 3 Type 4	0	1	1	12			.1	.43
18	Type 5	0	1	1	11			.3	.26
19	Type 6	0	1	1	8				
20	QC Type 7	1	10	1	53	2.5	2.82	1.3	2.83
21	QC Type 8	1	10	1	58			.1	3.09
22	QC Type 9	1	10	1	54				
23	QC Type 10	1	10	1	39			.2	3.38
24	Type 11	1	12	1	63			.3	3.48

[1]Lower Bound; [2]Upper Bound; [3]Weight; [4]Number in large pool;
[5]Per cent of a typical group with violation;
[6]Average number of items in a CAT for typical group.

Table 3: Content Constraint Violations for the Adaptive Quantitative Test, Smaller Pool, Extended Sympson-Hetter Method

Number	Description	LB[1]	UB[2]	W[3]	N[4]	Per cent of typical group	Ave number of items
1	Data Interp Set	2	2	11	19		
2	QC 1	5	5	10	39		
3	QC 2	4	4	10	34		
4	QC 3	4	4	10	34		
5	PS 1	3	3	10	26		
6	PS 2	3	3	10	26		
7	PS 3	2	2	10	18		
8	DI 1	4	4	1	135		
9	Type 4	8	8	1	178		
10	QC Type 4	2	2	10	20		
11	PS Type 4	2	2	10	23		
12	QC 1 Type 4	0	1	1	6	.1	.40
13	QC 2 Type 4	0	1	1	6	4.0	.70
14	QC 3 Type 4	0	1	1	8	9.0	.90
15	PS 1 Type 4	0	1	1	9	1.2	.48
16	PS 2 Type 4	0	1	1	9	14.9	.97
17	PS 3 Type 4	0	1	1	5	.1	.55
18	Type 5	0	1	1	8	.1	.20
19	Type 6	0	1	1	5		
20	QC Type 7	1	10	1	27	1.7	2.83
21	QC Type 8	1	10	1	27	.4	3.12
22	QC Type 9	1	10	1	28	.2	3.80
23	QC Type 10	1	10	1	25	.6	3.25
24	Type 11	1	12	1	49	.2	3.55

[1]Lower Bound; [2] Upper Bound; [3]Weight; [4]Number in smaller pool.

Table 4: Content Constraints and Weights for the Adaptive Verbal Test

Number	Description	LB[1]	UB[2]	W[3]	N(407)[4]	N(372)[5]
1	S:RCMP 1	1	1	10	13	13
2	S:RCMP 2	2	2	10	11	19
3	S:RCMP 3	1	1	10	10	11
4	S:RCMP 4	0	1	5	2	3
5	S:RCMP 5	0	1	5	3	5
6	S:RCMP 6	1	1	10	9	10
7	RCMP Items	8	8	10	153	184
8	RCMP 1	1	4	1	23	29
9	RCMP 2	1	4	1	37	43
10	RCMP 3	1	4	1	49	54
11	RCMP 4	1	4	1	32	40
12	RCMP 5	0	4	10	13	21
13	RCMP 6	0	4	10	21	32
14	SNCP	5	5	10	62	42
15	SNCP 1	0	2	1	14	11
16	SNCP 2	0	2	1	18	11
17	SNCP 3	0	2	1	16	11
18	SNCP 4	0	2	1	14	9
19	SNCP 5	0	1	2	5	4
20	SNCP 6	0	1	2	5	4
21	ANALOGIES	6	6	10	72	48
22	ANALOGIES 1	0	2	1	10	8
23	ANALOGIES 2	0	2	1	26	14
24	ANALOGIES 3	0	2	1	18	12
25	ANALOGIES 4	0	2	1	18	14
26	ANTM	8	8	10	96	66
27	ANTM 1	0	3	1	17	15
28	ANTM 2	0	3	1	28	18
29	ANTM 3	0	3	1	22	12
30	ANTM 4	0	3	1	29	21
31	Type 1	2	19	1	68	56
32	Type 2	2	19	1	80	69
33	Type 3	2	19	1	71	69
34	Type 4	2	19	1	80	68
35	Type 5	2	19	1	84	78

[1]Lower Bound; [2]Upper Bound; [3]Weight;
[4]Number in inadequate pool; [5]Number in adequate pool

Table 5: Content Constraint Violations for the Adaptive Verbal Test, Both Pools, ESH Exposure Control Method

Number	Description	LB[1]	UB[2]	W[3]	N[4]	Inadequate Pool Per cent[5]	Inadequate Pool, Ave n[6]	N[7]	Adequate Pool, Per cent[5]	Adequate Pool, Ave n[6]
1	S:RCMP 1	1	1	10	13			13		
2	S:RCMP 2	2	2	10	11			19		
3	S:RCMP 3	1	1	10	10	31.1	1.31	11		
4	S:RCMP 4	0	1	5	2			3		
5	S:RCMP 5	0	1	5	3			5		
6	S:RCMP 6	1	1	10	9	38.3	1.38	10		
7	RCMP Items	8	8	10	153			184		
8	RCMP 1	1	4	1	23	2.2	1.15	29	14.1	1.22
9	RCMP 2	1	4	1	37	.8	2.23	43	3.4	1.84
10	RCMP 3	1	4	1	49	3.2	2.97	54	5.8	2.27
11	RCMP 4	1	4	1	32	21.3	1.35	40	1.0	2.01
12	RCMP 5	0	4	10	13			21		
13	RCMP 6	0	4	10	21			32	1.0	1.26
14	SNCP	5	5	10	62			42		
15	SNCP 1	0	2	1	14	1.7	1.18	11	1.9	1.21
16	SNCP 2	0	2	1	18	.1	.98	11	.1	.92
17	SNCP 3	0	2	1	16	1.9	1.17	11	2.9	1.22
18	SNCP 4	0	2	1	14	5.9	1.66	9	6.4	1.65
19	SNCP 5	0	1	2	5	2.6	.56	4	.1	.51
20	SNCP 6	0	1	2	5			4		
21	ANALOGIES	6	6	10	72			48		
22	ANALOGIES 1	0	2	1	10	.5	1.00	8	.4	.93
23	ANALOGIES 2	0	2	1	26	13.6	1.92	14	17.0	1.96
24	ANALOGIES 3	0	2	1	18	.2	1.30	12		
25	ANALOGIES 4	0	2	1	18	15.4	1.79	14	13.8	1.78
26	ANTM	8	8	10	96			66		
27	ANTM 1	0	3	1	17	.5	1.60	15	.6	1.69
28	ANTM 2	0	3	1	28	3.0	2.21	18	3.5	2.23
29	ANTM 3	0	3	1	22	.1	1.36	12		
30	ANTM 4	0	3	1	29	20.4	2.83	21	1.3	2.77
31	Type 1	2	19	1	68	.9	4.51	56		
32	Type 2	2	19	1	80	.1	5.20	69	.3	4.59
33	Type 3	2	19	1	71	.9	5.49	69	.4	5.59
34	Type 4	2	19	1	80	1.4	4.94	68	.4	4.91
35	Type 5	2	19	1	84			78		

[1]Lower Bound; [2] Upper Bound; [3]Weight; [4]Number in inadquate pool; [5]Per cent of a typical group with violation; [6]Average number of items in a CAT for typical group; [7]Number in adequate pool.

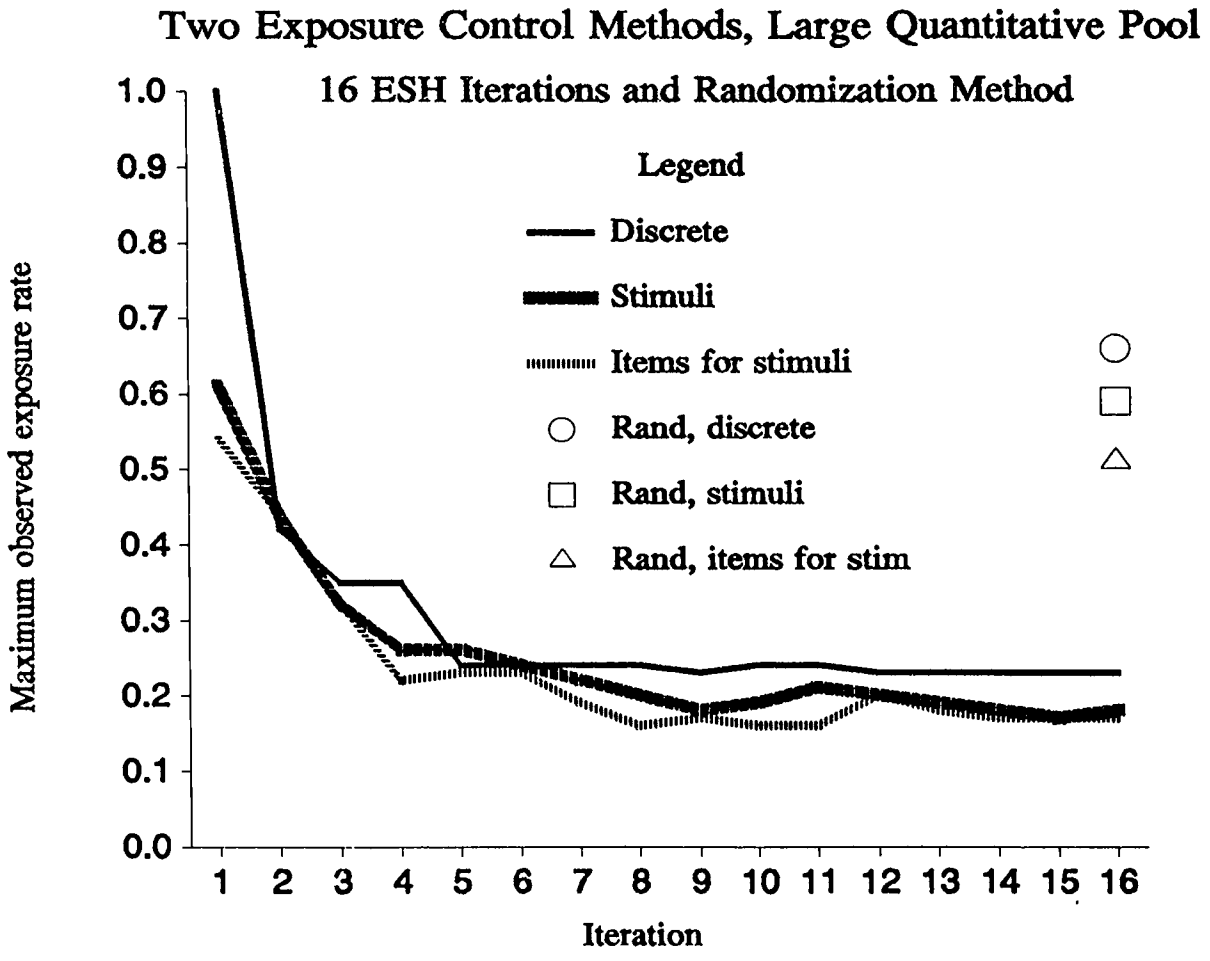

Figure 1: Two exposure control methods (16 extended Sympson-Hetter iterations and the randomization method) for a 25-item adaptive test and the large quantitative pool. See text.

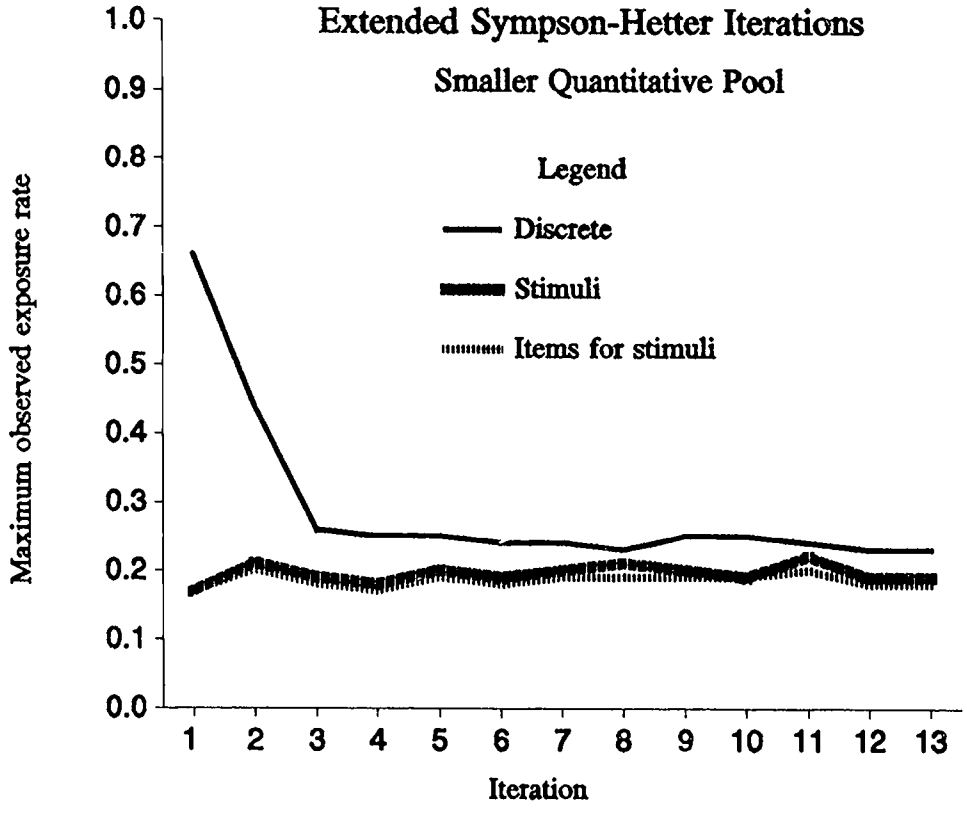

Figure 2: Iterations to obtain final exposure control parameters for a 25-item adaptive test and the smaller quantitative pool. See text.

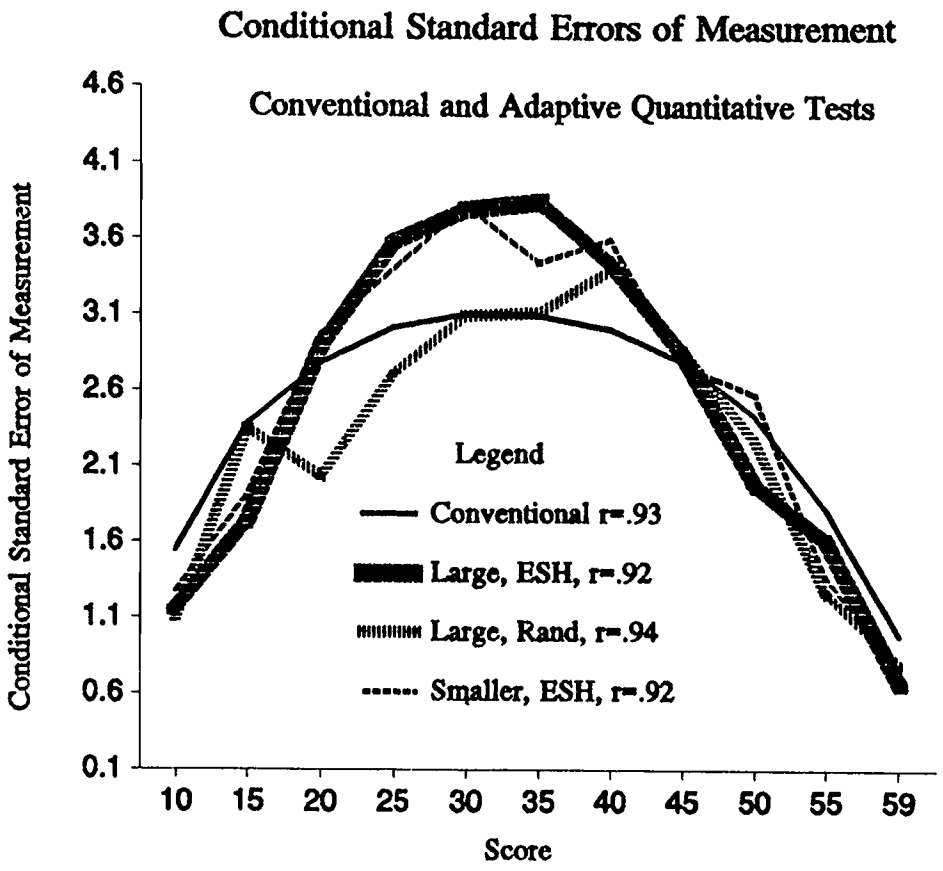

Figure 3: Conditional standard errors of measurement for a conventional test and three 25-item adaptive tests using two different pools and two different exposure control methods.

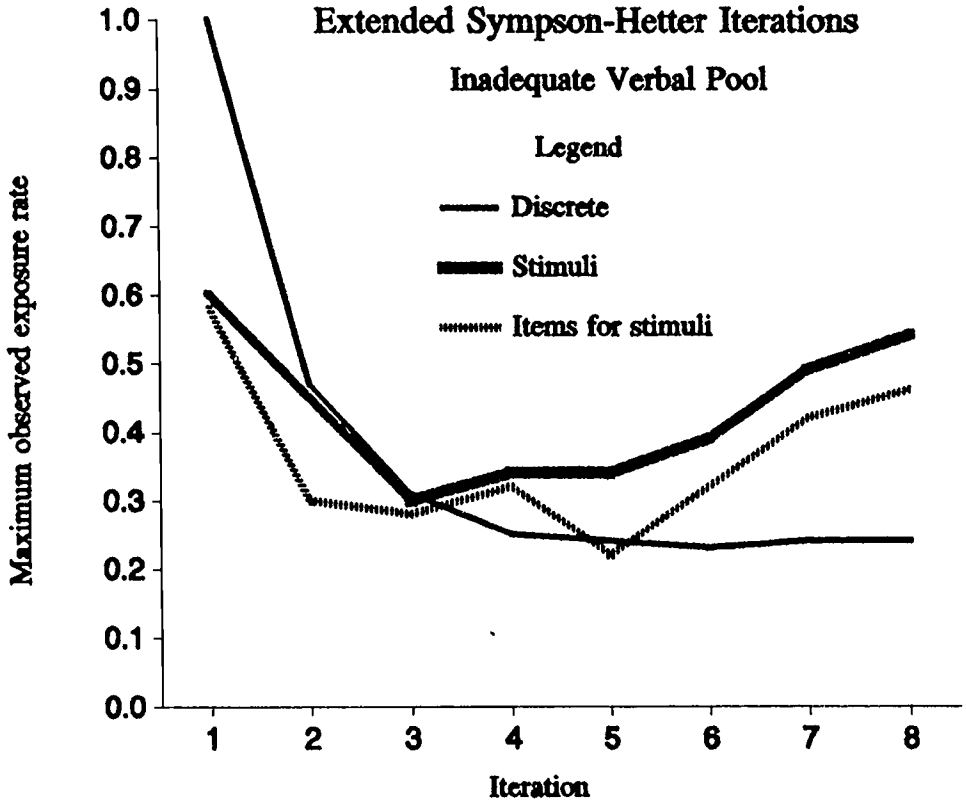

Figure 4: Iterations to obtain final exposure control parameters for a 27-item adaptive test and the inadequate verbal pool. See text.

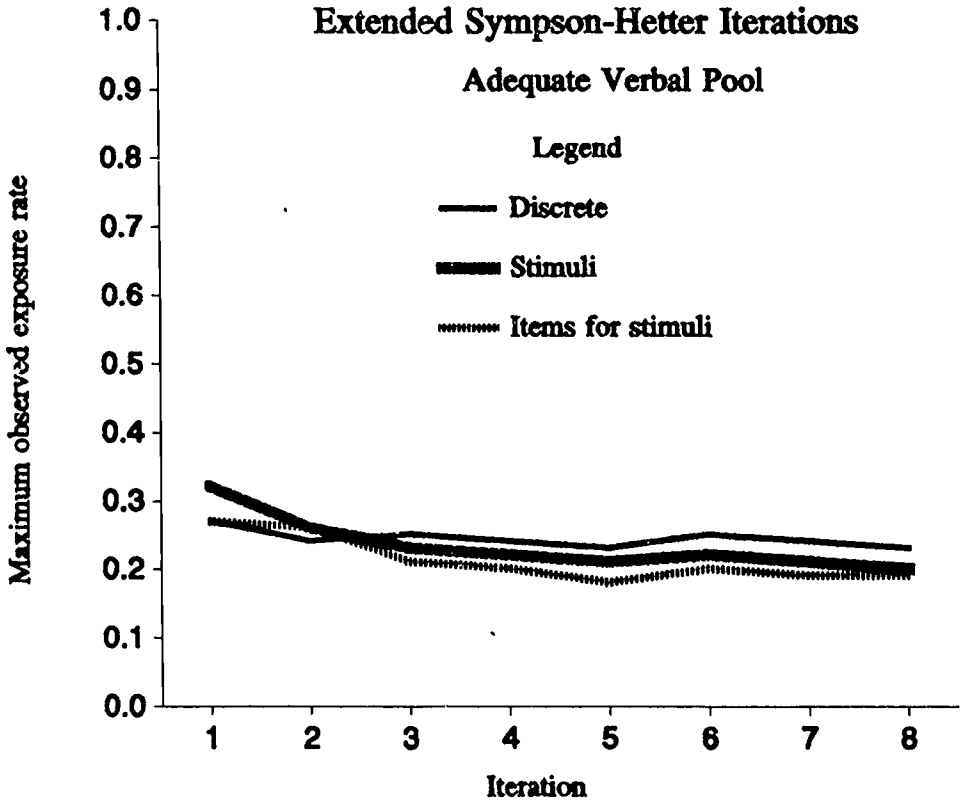

Figure 5: Iterations to obtain final exposure control parameters for a 27-item adaptive test and the adequate verbal pool. See text.